**연구노트**

**(수학과)이 재 봉**  2017.05.24 (목)

|  |  |  |
| --- | --- | --- |
|  | | |
| **연 구**  **주 제** | **특허분석을 통한 미래기술 예측** | |
| **자연언어 처리 기술** | |
| **구 분** | **연구내용** | **진행률** |
| **연 구**  **내 용** | [단어가방법]  초기의 몇 가지 텍스트마이닝 응용 분야 중에는 텍스트 기반의 문서집합에 구조를 도입할때 이들을 2개 이상의 미리 정해진 부류로 분류하거나 자연스런 군집으로 만들기 위해서 단어가방(bag-of-words)이라고 하는 단순화된 표현방법을 사용하였다. 단어가방 모형에서는 문장, 문단 또는 전체 문서 등과 같은 텍스트를 단어의 집합으로 푷ㄴ하는데, 문법이나 그 단어들이 나타내는 순서 등은 무시한다. 단어가방 모형은 아직도 일부 간단한 문서분류 도구에서 사용되고 있다. 예를 들어, 스펨메일 걸러내기에서 이메일 메시지를 단순히 순서 없는 단어들의 집합(단어가방)으로 모형화하고 미리 정해진 2개의 가방과 비교한다. 가방 하나는 스팸 메시지에서 발견되는 단어들로 채워진 것이고 다른 가방은 보통의 이메일에서 볼 수 있는 단어들이 들어 있다. 일부 단어들은 2개의 가방에 모두 들어 있지만 사용자의 친구나 일터에 관련이 있는 단어들을 더 많이 포함하고 있을 보통의 가방에 비해 ‘스팸’ 가방에서는 주식시장, 비아그라, 구입 등과 같이 스팸과 관련이 있는 단어들을 훨씬 더 많이 가지고 있을 것이다. 특정 이메일의 단어가방이, 서술자들을 포함한 두 가방들과 얼마만큼 일치 하는가로 스팸 여부를 결정한다.  하지만 원래 우리 인간은 단어들을 순서나 구문 없이 사용하지 않는다. 단어들은 통사론적 구조(syntactic structure)와 더불어 의미론적 구조(semantic structure)도 가지고 있는 문장 안에서 사용한다. 그러므로 자동화된 기술(텍스트 마이닝)이라면 단어가방 식의 해석을 넘어 보다 많은 의미론적 구조를 포함할 수 있는 방법을 찾아야만 한다. 텍스트 마이닝에서 현재의 경향은 자연언어 처리를 사용하여 얻을 수 있는 많은 향상된 특성들을 포함시키는 것이다.  단어가방방법은 텍스트마이닝을 할 때(예 : 분류, 군집화, 연관) 그다지 좋은 정보내용을 만들어내지 못할 수도 있다고 알려졌다. 그 좋은 예는 증거기반 의학에서 찾을 수 있다. 증거기반 의학의 핵심은 의료의사 결정 과정에 가장 좋은 가용한 연구결과들을 활용하는데, 출판매체로부터 수집된 정보들의 타당성과 적합성을 평가하는 것을 포함한다. 메릴랜드 대학교 연구진은 단어가방법을 사용해서 증거평가 모형을 개발하였다(Lin과 Demmer-Fushman, 2005). 이들은 기계학습법을 사용하여 MEDLINE(Medical Literature Analysis and Retrieval System)에서 50만개 이상의 연구논문들을 수집하였다. 이 모형은 각 초록들을 단어가방으로 나타내었는데 어간으로 만든 각각의 용어들을 특성으로 표현하였다. 검증된 실험계획법과 함께 잘 알려진 분류방법을 사용했지만 그들의 예측결과는 단순한 추측보다 나을게 없었으며, 이로써 단어가방법은 이 분야에서 연구논문들에 대해 충분히 좋은 결과를 만들어내지 못함을 알게 되었다. 따라서 자연언어 처리(NLP)와 같은 보다 진보된 기법이 필요하게 되었다.  [자연언어 처리(NLP)]  NLP는 텍스트마이닝의 중요한 요소이며, 인공지능과 컴퓨터언어학의 하부 분야이다. 이것은 컴퓨터 프로그램이 텍스트로 된 문서와 같은 인간 언어의 표현물을 다루기 쉽도록 형식을 더 갖춘, 즉 숫자와 기호의 형태를 가진 표현물로 변환하겠다는 목표 아래 인간의 자연언어를 이해하는 문제를 연구한다. NLP의 목표는 구문론에 의한 텍스트 다루기를 벗어나서 상황과 함께 문법과 의미론적 제약 사항을 고려하며 자연언어를 진정으로 이해하고 처리하고자 한다.  이해(understanding)라는 단어의 정의와 범위는 NLP에서 중요한 논쟁 주제 중의 하나이다. 인간의 자연언어는 모호하며, 의미를 정말로 이해하는 데에는 단어, 문장 그리고 문단 안에 무엇이 있는가를 넘어서 주제에 대한 광범위한 지식을 필요로 한다. 컴퓨터가 관연 인간이 하는 것과 같은 방식으로, 그리고 동일한 정확도로 자연언어를 이해할 수 있을까? 아마 아닐 것이다! NLP는 단순한 단어 개수를 세던 시절로부터 많이 지나왔지만 인간의 자연언어를 정말로 이해하려면 가야 할 길이 훨씬 더 많이 남아 있다.  인공지능 학계에서는 오랫동안 텍스트를 자동적으로 읽고 이로부터 지식을 얻을 수 있는 알고리즘을 만들고자 꿈꾸어 왔다. 스탠포드 대학교의 NLP 연구실은 파싱한 텍스트에 학습 알고리즘을 적용해서 텍스트로부터 자동으로 개념 및 개념들 간의 상호 관계들을 찾아내는 방법을 개발해왔다. 그들의 알고리즘은 많은 양의 텍스트에 특별한 방법을 적용해서 전 세계로부터 자동적으로 수십만 가지의 지식들을 획득하고 이를 응요하여 매우 훌륭한, WordNet 저장소를 만들어낸다. WordNet은 영어단어들과 그것들의 뜻, 동의어들 그리고 동의어 집합들 간 다양한 의미론적 관계 등을 수작업으로 공들여 코딩해서 만든 데이터베이스이다. 이것은 NLP 으용을 위한 주된 자원이지만 수작업으로 개발하고 유지하기에는 매우 비싸다. 지식이 자동적으로 WordNet안으로 들어올 수 있다면 WordNet은 저렴한 비용으로 NLP를 위한 보다 훌륭하고도 종합적인 자원이 될 수 있다.  NLP와 WordNet의 장점이 잘 드러나는 분야는 고객관계관리(CRM)이다. 광범위하게는 CRM의 모굪는 고객의 실제 및 알려진 니즈를 더 잘 이해하고 이에 효과적으로 반응함으로써 고객가치를 극대화하는 것이다. CRM에서 NLP의 여향을 크게 받은 중요한 분야는 감성분석이다. 감성분석은 웹 게시 형태로 주어지는 고객의 의견들과 같이, 대용량의 텍스트 형태의 데이터 소스를 이용하여 특정 상품과 서비스에 대한 호감과 비호감 의견들을 찾아내는 기법이다.  NLP는 이전에는 인간에 의해서만 할 수 있었던 인간의 자연언어를 자동적으로 처리해주는 컴퓨터 프로그램을 통해 다양한 영역에서 다양한 작업에 성공적으로 적용되어 왔다. 다음은 이들 작업 가운데 가장 널리 쓰이는 것들이다. |  |
| **차 주**  **계 획** | ■ 계획  - 관련 책 조사 | |
| **이 슈** | ■ **해당사항 없음** | |

|  |  |
| --- | --- |
|  |  |